Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 15:32:55
REFERENCE NO: 263

Author Names & Affiliations

- Chris Hill - MIT

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Computational Earth Science and also (2) Advanced Cyberinfrastructure and Research Computing Strategy and Practice

Title of Submission

On The Broad Need To Catalyze Greater Investment In Smart and Interoperable CyberInfrastructure and Research Computing

Abstract (maximum ~200 words).

The next 25 years will almost certainly see an explosion in the application of research computing technologies to creating new knowledge and innovation. The drivers for this are a tipping point that is becoming apparent in many fields, driven by dramatically falling costs in sensor devices, increases in network enabled autonomy and huge improvements in the ability to create and operate large scale systems. This confluence is allowing fields across disciplines spanning social science to plasma fusion to gather more information, process more information and reason about more information. There is a potential disconnect between on the ground innovation in research enterprises, and the current approaches to investment in cyberinfrastructure and research computing. In particular the nation may be falling short in its investments in this area by as much as $2B/year. This could be impacting the pace of innovation and the return on investment around many areas of central interest to the National Science Foundation. A focussed activity that makes the case for growing these infrastructure investments could have an outsized impact. Central to this could be a more formal recognition that research computing now has similar status as physical facilities and libraries in the budget vocabulary of the Federal government. Much of the investment now has such a
broad, general role that supporting it through indirect cost pools and through novel partnerships deserves serious analysis.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

A fundamental research challenge (and opportunity) that faces computational Earth science and almost every other field is the exploding availability of autonomous sensor data. This is true in laboratories and in the field. The coming decade will see more of this, whether through crowd sourced data gathering, cost effective cube satellites or web monitoring (for example). Examples from Earth science include ocean monitoring activities that are able to gather vastly more information through multiple in-situ and remote autonomous observing platforms. These can provide remarkable detail about the ocean internal wave field globally, for example. This can be important for possibilities around underwater positioning and communication, with potential commercial, disaster response preparedness and defense value. It is also important for basic understanding of the Earth system and its role in sustaining life. The internal wave field is a key piece of the oceans overall energy budget and plays roles in how ecosystems that in turn maintain global ocean, carbon and nitrogen balances behave. The interactions and dynamical processes involved are governed by relatively well known equations, but they are non-linear and lack clear scale separation. In the coming years with the right cyberinfrastructure it is likely to be possible to advance understanding through the application of machine learning/big-data methods as more digital information becomes available. In that scenario individual researchers may start to reason using many petabyte and beyond volumes of data in conjunction with data transport and compute needs to match. This is one example of many in Earth science. Examples like this exist in many fields. Often the examples are highlighted and sponsored as domain enterprises, however they likely possess many similar infrastructure needs.

The national research computing infrastructure is probably not fully prepared for the potential growth both technically and fiscally. It is still very cumbersome to move large volumes of digital information together in arbitrary ways, the nations network infrastructure is strong in many
point ways but still often sub-optimal in end-to-end capabilities. Innovations in the public cloud space are driving interesting, positive new machine learning and "smart cyberinfrastructure" dynamics, but they are also creating non-negligible risks of monopolistic data mobility lock-in. These could lead to business overreach, price gouging and control of core assets that ultimately belong to the nation as a whole. Any comprehensive national strategy for addressing this ad-hoc explosion in data capture, sharing, analysis and active curation is somewhat behind the curve. There are point solutions being created, but the interconnection and interoperability among those solutions is weak. Partly this is an inevitable consequence of rapid innovation and progress, but in part it comes from the practices around how these point solutions are funded and managed. In contrast with libraries and physical facilities approaches to providing resources, there is no explicit carve out in the Federal and other sponsor cost recovery processes or thinking for the enabling of research computing. This creates useful and valuable competition, but also challenges the potential of seamless collaboration. Historically this made sense as this was a relatively niche activity. This is no longer the case when a broad view is taken. Almost every laboratory and research group in an enterprise now has a digital activity that could scale considerably with good outcomes. The days when cyberinfrastructure was synonymous exclusively with niche high-performance computing are rapidly becoming a thing of the past. Today and looking forward the infrastructure needs are more broad based for end-to-end network, cost-effective storage and edge distribution, engines for processing using machine learning/big-data strategies for both statistical and symbolic inference and a skilled workforce ecosystem to complement. Today research computing infrastructure really needs to be more pervasive whether in Earth science or more broadly. A deliberate review and revision of how to catalyze this through multiple avenues of investment and funding policy could have a major impact. It could lay the way forward for new multi-billion dollar annual investments to be made strategically and broadly by multiple stake holders. Such a review should capture the justification for this level of investment, explore the possibility for a rethinking of how such items are perceived in foundational documents like OMB A21 and look to leverage and catalyze synergies across private sector (that is a tremendous beneficiary of basic discover and innovation), defense and intelligence concerns, not-for-profit research enterprises, and public entities in the form of states and cities.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

The internal wave problem described previously can be used as an illustration, but the reality is that nearly the same needs arise across domains. For the internal wave problem some of the needed cyberinfrastructure pieces would include (1) a data distribution fabric that connects very large multi-petabye data stores and compute resources seamlessly, (2) a continuous global ocean array of thermistor chains providing real time vertical structure information via satellite links, (3) an ongoing program of monitoring cube sats and aircraft, (4) a deep network of computational and data resources available for processing with low friction, (5) an effective market place to ensure cost-effective availability of resources, (6) ongoing investments in statistical and closed form equation AI for the automated creation of reduced order relations between unobserved and observed quantities, (7) significant investments in very large, highly networked (internally and externally) computing platforms geared to multi-scale, multi-physics, multi-algorithm hierarchical models, (8) investment in tools and workforce skills to utilize and operate this infrastructure in a cost-effective manner. Each of these areas would need investment as point solutions, but importantly achieving wholistic interoperation between these pieces is significant too to maximize impact. In different fields some of the sensor and algorithm components would be different, but a large number of the needed elements are common and investments could be shared in some manner. At present all the pieces exist to some degree, but they are hard to work with as a whole and need to continue to grow in scale.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

A possibly interesting historical extra data point to consider comes from the 1993 Branscombe report that contains a figure putting NSF high-end computing budget alongside other Federal agency amounts. It appears to show a much higher relative investment in 1993, that appears to have stayed relatively flat since then. It is clear that technology has moved on and it makes sense that some areas of investment have become more commercial and operational. It is also clear that commercialization has been enormously beneficial and that NSF ongoing investments are having good impact beyond that. However, it seems that some of the shift in the way cyberinfrastructure has become pervasive and critical is still poorly appreciated. An exploration of the notion that in many cases cyberinfrastructure is at least as important, if not more important, as office space and therefore should be thought of in a different way could help move the conversation toward the investments that are probably needed.